HARQ Buffer Management: An Information-Theoretic View

Wonju Lee, *Student Member, IEEE*, Osvaldo Simeone, *Senior Member, IEEE*, Joonhyuk Kang, *Member, IEEE*, Sundeep Rangan, *Senior Member, IEEE*, and Petar Popovski, *Senior Member, IEEE*

Abstract

A key practical constraint on the design of Hybrid automatic repeat request (HARQ) schemes is the size of the on-chip buffer that is available at the receiver to store previously received packets. In fact, in modern wireless standards such as LTE and LTE-A, the HARQ buffer size is one of the main drivers of the modem area and power consumption. This has recently highlighted the importance of HARQ buffer management, that is, of the use of buffer-aware transmission schemes and of advanced compression policies for the storage of received data. This work investigates HARQ buffer management by leveraging information-theoretic achievability arguments based on random coding. Specifically, standard HARQ schemes, namely Type-I, Chase Combining and Incremental Redundancy, are first studied under the assumption of a finite-capacity HARQ buffer by considering both coded modulation, via Gaussian signaling, and Bit Interleaved Coded Modulation (BICM). The analysis sheds light on the impact of different compression strategies, namely the conventional compression log-likelihood ratios and the direct digitization of baseband signals, on the throughput. Then, coding strategies based on layered modulation and optimized coding blocklength are investigated, highlighting the benefits of HARQ buffer-aware transmission schemes. The optimization of baseband compression for multiple-antenna links is also studied, demonstrating the optimality of a transform coding approach.

W. Lee and J. Kang are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 305-701, South Korea (e-mail: wonjulee@kaist.ac.kr, jhkang@ee.kaist.ac.kr).

O. Simeone is with the Center for Wireless Communications and Signal Processing Research (CWCSPR), Department of Electrical and Computer Engineering, New Jersey Institute of Technology (NJIT), Newark, NJ 07102, USA (e-mail: osvaldo.simeone@njit.edu).

S. Rangan is with the New York University Wireless Center, Department of Electrical and Computer Engineering, Polytechnic Institute of New York University (NYU), Brooklyn, NY 11201, USA (e-mail: srangan@poly.edu).

P. Popovski is with the Antennas, Propagation, and Radio Networking (APNET), Department of Electronic Systems, Aalborg University, Aalborg, 9220, Denmark (e-mail: petarp@es.aau.dk).

I. Introduction

Hybrid automatic repeat request (HARQ) is an integral part of modern wireless communication standards such as LTE and LTE-A [1], [2]. HARQ enables reliable communication over time-varying fading channel by leveraging both forward error-correcting coding at the physical layer and automatic retransmissions at the data link/medium access layer based on binary ACK/NACK feedback on the reverse link. With HARQ, the receiver can store previously received packets for joint processing with the last received packet in order to enhance the decoding reliability. Three HARQ mechanisms are conventionally used, namely HARQ Type I (HARQ-TI), HARQ Chase Combining (HARQ-CC), and HARQ Incremental Redundancy (HARQ-IR) (see, e.g., [1]-[4]).

One of the key challenges in implementing HARQ is the need to store data from previously received packets on chip. In LTE and LTE-A, the HARQ buffer is in fact one of the main drivers of the overall modem area and power consumption, as well as a key determinant of the User Equipment (UE) category level [2], [5]. Placing the HARQ buffer off chip can also be challenging due to the large bandwidth requirements on the external memory interface. These problems are expected to become even more severe for the next-generation systems, e.g., based on mmWave technology [6], [7], due to the larger bandwidth and transmission rates.

The limitations in the HARQ buffer size dictated by the modem area and power consumption make the use of buffer-aware transmission strategies and of advanced compression[1] policies for the storage of received data of critical importance for the feasibility of HARQ in modern wireless standards [5], [8]. An example of the former is limited buffer rate matching in LTE [2] and an instance of the latter is the vector quantization scheme proposed in [8] to store the log-likelihood ratios (LLRs) of the coded bits for the previously received packets. We refer to transmit- and receive-side mechanisms meant to cope with HARQ buffer limitations as HARQ buffer management.

Previous theoretical work on HARQ has assumed unrestricted HARQ buffers to be available at the receivers or has imposed limits on the number of packets that can be stored (see, e.g., [3], [9] and references therein). In this paper, instead, we assume a generic capacity constraint for the HARQ buffer in terms of number of bits, and we aim at addressing the following main questions: (*i*) How is the relative performance of standard HARQ schemes, namely HARQ-TI, HARQ-CC and HARQ-IR, affected by the amount of available HARQ buffer capacity? (*ii*) Are there more efficient alternatives to the conventional approach of representing buffered packets at the receiver by quantizing the LLRs of the coded bits (see [5], [8])? (*iii*) What is the impact of buffer-aware transmission strategies such as layered modulation and rate matching? (*iv*) What new opportunities and challenges arise in the design of HARQ buffer management for multiple-antenna (MIMO) links?

[1]In this paper, compression is meant to include also the step of quantization.

This works makes some steps towards answering these questions by leveraging information-theoretic achievability arguments based on random coding. Our contributions are as follows.

- We study a baseline system that uses an ideal coded modulation scheme via Gaussian signaling at the transmitter and compression of the previously received packets at the baseband level with the aim of assessing the impact of a finite HARQ buffer on the throughput of HARQ-TI, HARQ-CC and HARQ-IR (Sec. III).
- We investigate the more complex case of a link employing Bit Interleaved Coded Modulation (BICM) [10] and study the performance with both baseband compression and the more conventional LLR compression of the previously received packets (Sec. IV). The goal of the analysis is to address the possible suboptimality of the conventional approach of quantizing LLRs for storage in the HARQ buffer.
- We study the potential benefits of buffer-aware transmission strategies based on layered transmission [11], whereby the rates of the transmission layers are adopted to the HARQ buffer size (Sec. V).
- We study the design of baseband compression for a link with multiple-antennas and show the optimality of a compression strategy based on transform coding (Sec. VI).
- We analyze the impact of the selection of the transmission blocklength as a function of the HARQ buffer size (Sec. VII). This analysis complements the study in [12], which assumed no buffer limitations.

Finally, Sec. VIII presents numerical results and Sec. IX offers with some concluding remarks[2].

Notation: $(\cdot)^*$ denotes the complex transpose; $E[\cdot]$ is the expectation operator; information-theoretic quantities such as mutual information are defined as in [14].

II. System Model and Performance Criteria

Throughout this paper, except for Sec. VI, we consider a communication link with a single-antenna transmitter and a single-antenna receiver operating over a quasi-static fading channel via an HARQ mechanism. As illustrated in Fig. 1 and further discussed below, we make the assumption that the receiver has a limited HARQ buffer to store information extracted from the packets received in the previous (re)transmissions. Time is slotted and each slot accommodates the transmission of a packet of length L symbols. The received signal in a channel use of the i-th slot is given by

$$Y_i = \sqrt{\mathrm{SNR}} H_i X_i + Z_i, \tag{1}$$

where the parameter SNR represents the average signal to noise ratio; the channel gain H_i has unit power and changes independently slot by slot with a given cumulative distribution function (cdf) F; the input signal X_i is subject to the power

[2]The content of Sec. III and Sec. IV was partially presented in [13].

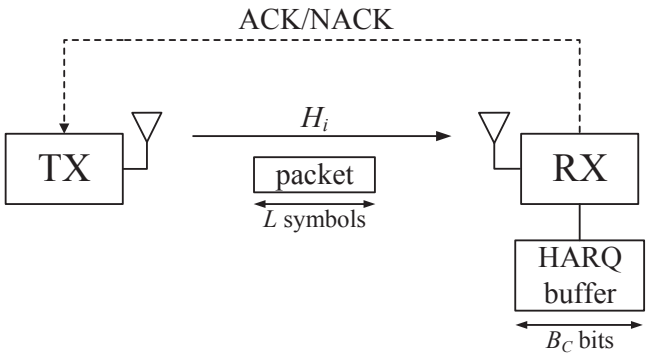

Fig. 1. HARQ with a limited-capacity HARQ buffer. Except for Sec. VII, we set $B_C = LC$, where C is the buffer size normalized to the packet length.

constraint $E[|X_i|^2] = 1$; and we have the additive noise $Z_i \sim \mathcal{CN}(0, 1)$. The receiver has an HARQ buffer with capacity B_C bits. Except for Sec. VII, we will set $B_C = LC$, where C is hence the buffer size normalized with respect to the packet length. The channel gain H_i is assumed to be known to the receiver, where, being a single (complex) value per packet, it is stored using a negligible buffer space.

Let us denote the maximum number of retransmission by N_{max} and the transmission rate by R, which is measured in bits/s/Hz or, equivalently, in bits/symbol. Note that, unless stated otherwise, we consider single-layer modulation at rate R. The case of multi-layer modulation will be considered in Sec. V. Moreover, except for Sec. VII, the blocklength L will be considered to be long enough so as to justify the use of information-theoretic asymptotic bounds. Each HARQ session, of at most N_{max} retransmission, including the original, hence aim at delivering a data packet of LR bits. For single layer modulation, the throughput T can be written as (see, e.g., [11])

$$T = \frac{R(1 - P_e^{N_{max}})}{E[N]}, \tag{2}$$

where N a random variable that measures the number of retransmissions, including the original transmission, which satisfies

$$E[N] = \sum_{n=1}^{N_{max}} n\text{Pr}[N = n]; \tag{3}$$

and P_e^n is the probability of an unsuccessful transmission up to, and including, the n-th attempt. We have

$$\text{Pr}[N = n] = P_e^{n-1} - P_e^n \tag{4}$$

for $n < N_{max}$ and $\text{Pr}[N = N_{max}] = P_e^{N_{max}-1}$. Therefore, from (2), it is sufficient to calculate the probabilities P_e^n for $n = 1, \ldots, N_{max}$ in order to characterize the throughput of any given HARQ scheme. We observe that (2) will need to be modified to account for layered modulation.

III. Gaussian signaling with baseband compression

In this section, we evaluate the throughput of HARQ-TI, HARQ-CC, and HARQ-IR assuming a baseline scheme whereby the transmitter uses Gaussian signaling and the receiver stores in the memory compressed version of the received baseband packets. Note that, in practice, Gaussian signaling can be interpreted as the use of an ideal coded modulation strategy at the transmitter (see, e.g., [9]).

A. HARQ-TI

With HARQ-TI, the transmitter repeatedly sends the same encoded packet and the receiver attempts decoding based solely on the last received packet. HARQ-TI hence does not make use of the receiver's HARQ buffer. Under the said assumption of sufficiently large L, the probability of an unsuccessful transmission up to the n-th attempt can be obtained as

$$\begin{aligned} P_e^n &= \Pr\left[\bigcap_{i=1}^{n}\left(\text{SNR}|H_i|^2 \leq 2^R-1\right)\right] \\ &= \prod_{i=1}^{n}\Pr\left[\text{SNR}|H_i|^2 \leq 2^R-1\right] = \left(F\left(\frac{2^R-1}{\text{SNR}}\right)\right)^n. \end{aligned} \tag{5}$$

We recall that the throughput is finally obtained as (2), which, in the case of HARQ-TI can be simplified as $T = R\left(1-P_e^1\right)$. Note that the throughput of HARQ-TI does not depend on N_{max}.

B. HARQ-CC

With HARQ-CC, the transmitter repeats the same packet at each retransmission as for HARQ-TI, but the receiver performs decoding on a packet obtained by combining all previously received packets via maximum ratio combining (MRC). HARQ-CC hence requires storage either of all previously received packets or of the current combined packet obtained from all previous transmissions. In the presence of a limited-buffer receiver, these two HARQ buffer management options yield different throughputs and are discussed next.

1) HARQ-CC Store and Combine (S&C): A first option to implement HARQ-CC in the presence of a limited HARQ buffer is for the receiver to store all the previously received packets. Due to memory limitations, prior to storage, packets need to be compressed. To this end, as illustrated in Fig. 2, the receiver divides the available memory size equally among all the packets received up to the given retransmissions and compresses each packet separately. If the n-th transmission is unsuccessful, the receiver then compresses the last received packet to LC/n bits and recompresses the previously stored packets to LC/n bits (from their previous larger size of $LC/(n-1)$ bits). We refer to this scheme as Store and Combine (S&C).

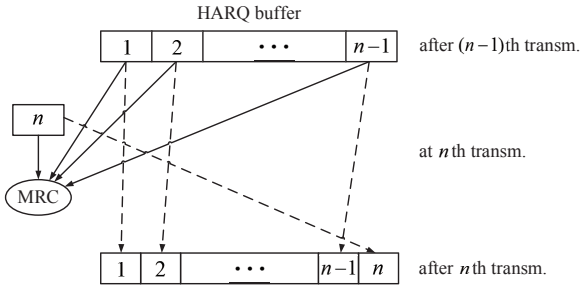

Fig. 2. Illustration of HARQ-CC S&C. The numbers indicate the index of the packet, which is compressed in the HARQ buffer, and the dashed lines correspond to compression operations that take place in case the n-th transmission fails. For the packets within the HARQ buffer recompression is carried out according to the successive refinement scheme discussed in Appendix.

In order to account for the effect of quantization, we use the standard additive quantization noise model. Specifically, if the n-th retransmission is not successful, the quantized signals are given by

$$\hat{Y}_{i,n} = Y_i + Q_{i,n}, \tag{6}$$

for $i = 1, \ldots, n$ and $n = 1, \ldots, N_{max}$, where $Q_{i,n} \sim \mathcal{CN}(0, \sigma^2_{i,n})$ is the quantization noise for the i-th received packet as stored at the n-th unsuccessful transmission. As discussed below, the quantization noise $\sigma^2_{i,n}$, which corresponds also to the mean squared error distortion, is adjusted to the current channel realization H_i, and hence quantization must be performed after channel estimation.

Remark 1. *Quantization noise models such as (6) are used throughout this work within the information-theoretic framework of random coding, and hence the quantization noise distribution is to be considered as obtained by averaging over the randomly selected vector quantization codebooks (see, e.g., [14], [15]). Moreover, following Shannon's classical arguments, the results obtained in this paper are to be interpreted as implying the existence of specific (deterministic) coding and compression strategies that achieve the calculated throughput levels as long as they operate over sufficiently long block-lengths [14], [15] (see Sec. VII for further discussion). From a practical viewpoint, results in [16] and [17] suggest that high-dimensional lattice vector quantizers, such as standard Trellis Coded Quantization (TCQ) [18], or graphical codes with message passing are expected to perform close to the performance evaluated in this work. However, the choice of a Gaussian distribution for the quantization noise is made with no claim of optimality and may be in practice justified by the fact that dithered lattice vector quantizers are able to approximate (6) with increasing accuracy as the dimensions of the quantizer increases [16]. Moreover, the Gaussian assumption implies that the performance evaluated here with baseband compression can be realized by receivers that use conventional minimum-distance decoders [19].*

Following Remark 1, we relate the quantization noise $\sigma^2_{i,n}$ to the number of allocated bits LC/n via the standard rate-

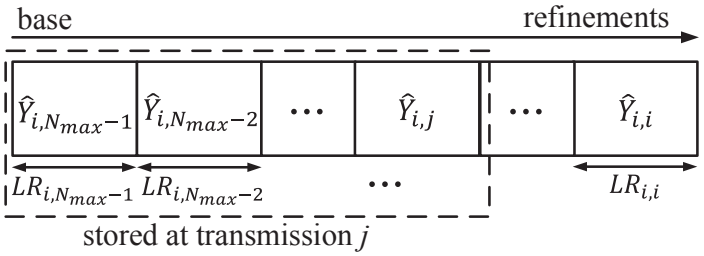

Fig. 3. Illustration of the successive refinement compression strategy used for HARQ-CC S&C and HARQ-IR: Each packet i is compressed to $(N_{max}-i)$ layers such that higher layers, corresponding to refinements, are discarded as n increases in order to free memory space for the more recent packets (see, e.g. Fig. 2).

distortion theoretic equality [14] $C/n = I(Y_i;\hat{Y}_{i,n})$, which can be evaluated as

$$\frac{C}{n} = \log_2\left(1+\frac{\text{SNR}|H_i|^2+1}{\sigma_{i,n}^2}\right) \tag{7}$$

implying

$$\sigma_{i,n}^2 = \frac{\text{SNR}|H_i|^2+1}{2^{C/n}-1}. \tag{8}$$

The equality (7) holds also for recompressed packets, i.e. for all packets (1) with $i<n$, as long as successive refinement compression [14, Ch. 13] is employed. Specifically, as illustrated in Fig 3. each packet i is first compressed at the i-th transmission (if unsuccessful) with a number $(N_{max}-i)$ of compression layers. At later transmissions, higher layers, corresponding to refinement descriptions, are progressively discarded as n increases in order to satisfy the rate constraint C/n and effectively increasing the quantization noise (8). We refer to Appendix for a detailed discussion.

At the n-th retransmission, the decoder performs MRC of the stored $(n-1)$ packets and of the last received packet prior to decoding as

$$\bar{Y}_n = H_n^* Y_n + \sum_{i=1}^{n-1} H_i^* \hat{Y}_i. \tag{9}$$

As a result, the effective SNR is equal to

$$\frac{\text{SNR}\left(\sum_{i=1}^{n}|H_i|^2\right)^2}{|H_n|^2+\sum_{i=1}^{n-1}|H_i|^2\left(\sigma_{i,n}^2+1\right)}, \tag{10}$$

and the probability of an unsuccessful transmission up to the n-th attempt is given by

$$P_e^n = \text{Pr}\left[\bigcap_{j=1}^{n}\left(\log_2\left(1+\frac{\text{SNR}\left(\sum_{i=1}^{j}|H_i|^2\right)^2}{|H_j|^2+\sum_{i=1}^{j-1}|H_i|^2\left(\sigma_{i,j}^2+1\right)}\right)\le R\right)\right]. \tag{11}$$

The probability in (11) can be calculated via Monte Carlo simulations and the same will apply also to the other probabilities indicated in the rest of the paper.

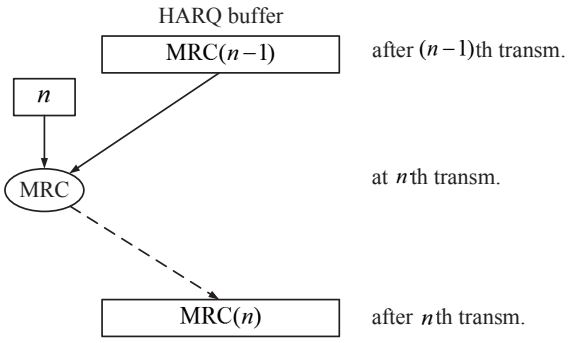

Fig. 4. Illustration of HARQ-CC C&S. MRC(n) indicates the compressed MRC-combined packet stored at the n-th transmission if unsuccessful.

Remark 2. *In the absence of memory restrictions, i.e., with $C \to \infty$, we have $P_e^n = Pr\left[\sum_{i=1}^n SNR|H_i|^2 \leq 2^R - 1\right]$ since $\sigma_{i,j}^2 \to 0$, hence obtaining the standard performance of Chase combining (see, e.g., [11]). Therefore, under this conventional assumption, there is no need to include the intersection operation in (11). This is because, with $C \to \infty$, the effective SNR (i.e., the ratio in (11)) is a monotonically increasing function of n, while this is generally not the case for finite C due to the increasing quantization noise power (8).*

Remark 3. *The combining (9) does not account for the different noise powers affecting the combined packets due to the quantization noise. Therefore, the combining (9) is suboptimal for finite C in terms of the achievable rate. In fact, it reflects the operation of a standard Chase combiner [4], which is oblivious to the presence of quantization effects. For reference, we observe that an optimal combining would achieve an effective SNR of (see, e.g. [11])*

$$SNR|H_n|^2 + \sum_{i=1}^{n-1} \frac{SNR|H_i|^2}{1+\sigma_{i,n-1}^2}, \tag{12}$$

which is generally larger than (10) and it coincides with (10) for $C \to \infty$. We also refer to Sec. III-D for a discussion of an adaptive storing scheme that uses the standard Chase combiner but decides adaptively whether to store a packet or not depending on the effective SNR improvement obtained as a result.

2) HARQ-CC Combine and Store (C&S): The S&C approach is expected to be inefficient since decoding is carried out on the combined packet and not on the individual packets, which thus need not be separately stored. For this reason, here, rather than storing all the previously received packets as with S&C, we consider compressing and storing directly the MRC-combined packet. Specifically, as illustrated in Fig. 4, at each retransmission, the last received packet is combined with the current stored packet in the HARQ buffer. If decoding is unsuccessful, the combined packet is compressed and stored. We refer to this scheme as Combine and Store (C&S).

To elaborate, if decoding is not successful at the first transmission, the compressed packet is given by

$$\begin{aligned}\hat{Y}_1 &= H_1^* Y_1 + Q_1 \\ &= \sqrt{\text{SNR}}|H_1|^2 X + H_1^* Z_1 + Q_1 \\ &= \sqrt{\text{SNR}}|H_1|^2 X + E_1, \end{aligned} \tag{13}$$

where $Q_1 \sim \mathcal{CN}\left(0, \sigma_1^2\right)$ is the quantization noise and $E_1 = H_1^* Z_1 + Q_1 \sim \mathcal{CN}\left(0, \rho_1^2\right)$ is referred to as the effective noise. From rate-distortion theory, similar to (8), we have $\sigma_1^2 = \left(|H_1|^2 + \text{SNR}|H_1|^4\right) / \left(2^C - 1\right)$ and $\rho_1^2 = |H_1|^2 + \sigma_1^2$. The combined signal used in decoding at the n-th transmission is given by

$$\bar{Y}_n = H_n^* Y_n + \hat{Y}_{n-1}, \tag{14}$$

for all $n > 1$. Moreover, the stored packet at the n-th attempt, if is unsuccessful, can be written as

$$\begin{aligned}\hat{Y}_n &= \bar{Y}_n + Q_n \\ &= \sqrt{\text{SNR}} \sum_{i=1}^{n} |H_i|^2 X + E_n, \end{aligned} \tag{15}$$

with the effective noise given by $E_n = E_{n-1} + H_n^* Z_n + Q_n \sim \mathcal{CN}\left(0, \rho_n^2\right)$. The power of the effective noise can be expressed using the recursive relationship

$$\rho_n^2 = \rho_{n-1}^2 + |H_n|^2 + \left\{ \rho_{n-1}^2 + |H_n|^2 + \text{SNR} \left(\sum_{i=1}^{n} |H_i|^2 \right)^2 \right\} \Big/ \left(2^C - 1\right). \tag{16}$$

Based on (14) and (16), we can finally obtain the probability of an unsuccessful transmission up to the n-th attempt as

$$P_e^n = \Pr\left[\bigcap_{j=1}^{n} \left(\log_2 \left(1 + \frac{\text{SNR}\left(\sum_{i=1}^{j} |H_i|^2 \right)^2}{|H_j|^2 + \rho_{j-1}^2} \right) \leq R \right) \right], \tag{17}$$

where we set $\rho_0 = 0$.

Remark 4. *As $C \to \infty$, the effective noise is given by $\rho_n^2 = \sum_{i=1}^{n} |H_i|^2$ and we have $P_e^n = Pr\left[\sum_{i=1}^{n} SNR|H_i|^2 \leq 2^R - 1\right]$. The other considerations made in Remark 2 and Remark 3 apply here as well.*

Remark 5. *As per Remark 3, the MRC operation in (15) neglects the fact that the noise power on the received signal at the current n-th transmission is different from the effective noise power ρ_{n-1}^2 that affects the previously received and combined packets due to the quantization noise. It hence reflects the operation of a standard Chase combiner [4].*

C. HARQ-IR

With HARQ-IR, at each retransmission, the transmitter sends a packet consisting of new parity bits from a rate-compatible code and decoding is based on the concatenation of all previously received packets. We assume that the receiver stores all the

previously received packets by following the same mechanism as in HARQ-CC S&C, and hence allocating the same buffer space to all packets. Note that the idea of storing a combined version of the previous baseband packets as in HARQ-CC is not directly applicable to HARQ-IR. The compressed packets at the n-th retransmission are given by (6) and (8). Since with HARQ-IR the achievable rate is the sum of the achievable rates across all transmissions (see, e.g. [9]), the probability of an unsuccessful transmission up to the n-th attempt can be obtained as

$$P_e^n = \Pr\left[\bigcap_{j=1}^{n}\left(\log_2\left(1+\mathrm{SNR}|H_j|^2\right)+\sum_{i=1}^{j-1}\log_2\left(1+\frac{\mathrm{SNR}|H_i|^2}{1+\left(1+\mathrm{SNR}|H_i|^2\right)/\left(2^{C/(j-1)}-1\right)}\right)\leq R\right)\right]. \tag{18}$$

Remark 6. *With $C\to\infty$, we obtain $P_e^n = Pr\left[\sum_{i=1}^{n}\log_2\left(1+SNR|H_i|^2\right)\leq R\right]$ [9] (see also Remark 2 and Remark 4). Moreover, by setting $C\to 0$ and $N_{max}=1$, the HARQ-IR throughput tends to that of HARQ-TI as it can be seen by comparing (18) and (5).*

D. Adaptive Storing

So far, we have assumed that all received packets are stored either individually or after MRC. However, in order to avoid using the available HARQ buffer capacity for received packets that do not carry significant information, one could instead store a packet only if the achievable rate is sufficiently improved as a result. This is particulary significant since, as discussed in Remarks 3, 4 and 6, the rate achievable with the studied conventional HARQ schemes does not necessarily increase with the number of retransmissions n. Here, we propose an *adaptive storing* strategy that is motivated by these observations.

We first describe the idea for HARQ-CC S&C. Let us define a random variable $N_s(n)$ that accounts for the number of packets that have been stored prior to transmission $n+1$. At transmission n, we first check if the achievable rate in the left-hand side of the inequality in (11) is larger than $\eta\geq 1$ times the achievable rate for the previous $n-1$ transmissions, where η is a design parameter. If so, the packet is stored and we set $N_s(n+1)=N_s(n)+1$; if not, the packet is not stored and $N_s(n+1)=N_s(n)$. To evaluate the performance of this scheme, in (11), the sums are restricted only to the indices of the $N_s(n)$ stored packets. Note that $N_s(n)$ and the indices of the stored packets are functions of the channel gains $|H_i|^2$ for $i=1,\ldots,n-1$. For HARQ-CC C&S and HARQ-IR, adaptive storing can be implemented and analyzed by following the same procedure described above, using the achievable rate appearing in the left-hand side of the inequality in (17) and the rate expression on the left-hand side of (18), respectively, in lieu of (11).

IV. BICM WITH BASEBAND AND LLR COMPRESSION

In this section, we consider transmission based on BICM with a fixed M-ary constellation $\mathcal{X}$, where $M=2^m$ for some integer m [10]. The main motivation for this investigation, beside the practical relevance of BICM, is the aim of conforming

baseband compression, as studied in the previous section, with a more conventional implementation whereby the receiver compresses the LLRs of the coded bits in the previously received packets (see, e.g., [5]). It is recalled that BICM maps coded bits directly on constellation points, hence facilitating the implementation and analysis of LLR processing and enabling the study of the impact of the constellation size.

Throughout this section, we make the standard assumptions of ideal interleaving, so that the m bit channels can be assumed to be independent, of a binary i.i.d. Ber$(1/2)$ codewords transmitted across all bit channels and of Gray mapping [10]. To elaborate, we define the j-th bit in the binary label of $X \in \mathcal{X},\ j = 1, \cdots, m$, according Gray mapping, as $X(j)$, and the set

$$\mathcal{X}_b^j = \left\{ x \in X \middle| X(j) = b \right\}, \tag{19}$$

for $b \in \{0, 1\}$, of all constellation points in which the j-th bit $X(j)$ equals b. With these definitions and (1), the LLR for the j-th bit of a symbol within the i-th retransmitted packet can be written as

$$L_i^j = \log_2 \frac{\sum_{x \in \mathcal{X}_1^j} \exp\left(- \left| Y_i - \sqrt{\text{SNR}} H_i x \right|^2 \right)}{\sum_{x \in \mathcal{X}_0^j} \exp\left(- \left| Y_i - \sqrt{\text{SNR}} H_i x \right|^2 \right)}. \tag{20}$$

In the rest of this section, we first review the performance of HARQ-TI, which does not require the use of the HARQ buffer and then study the performance of HARQ-CC and HARQ-IR first with baseband compression and then with LLR compression.

A. HARQ-TI

In order to evaluate the achievable rates with BICM, we first introduce the conditional probability density function (pdf) of Y_i given the j-th bit $X_i(j)$, which, from (1), is given by

$$f_{Y_i|X_i(j)}(y|b) = \frac{1}{2^{m-1}} \sum_{x \in \mathcal{X}_b^j} \frac{1}{\pi} \exp(-|Y_i - \sqrt{\text{SNR}} H_i x|^2), \tag{21}$$

using the fact that all binary variables $X_i(j)$ are i.i.d. Ber$(1/2)$. Due to joint encoding across the m bit channels, an outage event takes place when the m bit channels together do not support the transmission rate. Therefore, with HARQ-TI, the probability of an unsuccessful transmission up to the n-th retransmission can then be calculated as (see, e.g., [20])

$$\begin{aligned} P_e^n &= \prod_{i=1}^{n} \Pr\left[\sum_{j=1}^{m} I(X_i(j); Y_i) \le R \right] \\ &= \prod_{i=1}^{n} \Pr\left[\frac{1}{2} \sum_{b=0}^{1} \sum_{j=1}^{m} \int f_{Y_i|X_i(j)}(y|b) \log_2 \frac{f_{Y_i|X_i(j)}(y|b)}{f_{Y_i}(y)} dy \le R \right], \end{aligned} \tag{22}$$

with $f_{Y_i}(y) = 1/2 \sum_{b=0}^{1} f_{Y_i|X_i(j)}(y|b)$, where the second equality follows by direct calculation of the mutual information. While a closed-form expression for the conditional pdf $f_{Y_i|X_i(j)}(y|b)$ appears to be difficult to obtain, this quantity, and hence also (22), can be estimated numerically through Monte-Carlo simulations. Note that the throughput (2) can be simplified as $T = R(1 - P_e^1)$.

B. Baseband Compression

In this subsection, we consider the performance of HARQ-CC and HARQ-IR in the presence of baseband compression.

1) HARQ-CC: Similar to (21), in order to evaluate the performance of HARQ-CC, we first define the conditional pdf $f_{\bar{Y}_n|X(j)}(y|b)$ with $\bar{Y}_n$ being the combined packet, which is given by (9) for HARQ-CC S&C and (14) for HARQ-CC C&S. In particular, for HARQ-CC S&C, we obtain

$$f_{\bar{Y}_n|X(j)}(y|b) = \frac{1}{2^{m-1}} \sum_{x \in \mathcal{X}_b^j} f_{\bar{\mu}_n, \bar{\sigma}_n^2}(y), \tag{23}$$

where $f_{\bar{\mu}_n, \bar{\sigma}_n^2}(y) = 1/(\pi\bar{\sigma}_n^2)\exp(-|y-\bar{\mu}_n|^2/\bar{\sigma}_n^2)$ is the pdf of a complex Gaussian variable with mean $\bar{\mu}_n = \sqrt{\text{SNR}}\sum_{i=1}^{n}|H_i|^2 x$ and variance $\bar{\sigma}_n^2 = |H_n|^2 + \sum_{i=1}^{n-1}\text{SNR}|H_i|^2(\sigma_{i,n-1}^2+1)$ using (8), while for HARQ-CC C&S, we have the same mean $\bar{\mu}_n$ and variance $\bar{\sigma}_n^2 = |H_n|^2 + \rho_{n-1}^2$ with ρ_n^2 in (16). We can then write $P_e^n = \Pr\left[\sum_{j=1}^{m} I(X(j);\bar{Y}_i) \leq R\right]$, which can be calculated as (22).

2) HARQ-IR: Following similar arguments as for HARQ-CC and recalling Sec. III-C, the probability of an unsuccessful transmission up to the n-th attempt can be calculated as

$$P_e^n = \Pr\left[\bigcap_{i=1}^{n}\left(\sum_{j=1}^{m} I(X_n(j);Y_n) + \sum_{k=1}^{i-1}\sum_{j=1}^{m} I(X_k(j);\hat{Y}_{k,i}) \leq R\right)\right], \tag{24}$$

where the conditional pdf $f_{Y_i|X_i(j)}$ is given by (21) and the conditional pdf $f_{\hat{Y}_{k,i}|X_k(j)}$ of the compressed packet $\hat{Y}_{k,i}$ given by $X_k(j)$ is equal to $f_{\bar{\mu}_k, \bar{\sigma}_{k,i}^2}(y)$ with mean $\bar{\mu}_k = \sqrt{\text{SNR}}H_k x$ and variance $\bar{\sigma}_{k,i}^2 = \sigma_{k,i-1}^2 + 1$ in (8).

C. LLR Compression

Here we study the performance of HARQ-CC and HARQ-IR in the presence of LLR compression.

1) HARQ-CC: For HARQ-CC, as done in Sec. IV-B1, we consider both compression mechanisms S&C and C&S.

a) HARQ-CC Store and Combine (S&C): With LLR compression, similar to Sec. III-B, HARQ-CC S&C divides the available memory equally to store the compressed LLRs of the previous received packets for all bits channels. Specifically, at the n-th transmission, if unsuccessful, the compressed LLR for the i-th transmissions and bit channel j is given as

$$\hat{L}_{i,n}^j = L_i^j + Q_{i,n}^j, \tag{25}$$

for $i = 1, \ldots, n$ and $n = 1, \ldots, N_{max}$, where we follow the same standard additive quantization noise model used in Sec. III and the quantization noise is modelled as $Q_{i,n}^j \sim \mathcal{N}(0, \sigma_{i,n,j}^2)$ (see Remark 1 for a discussion on this model). To evaluate the quantization noise variance $\sigma_{i,n,j}^2$, we resort to the information-theoretic equality $I(L_i^j; \hat{L}_{i,n}^j) = C/(mn)$, which accounts for the fact that each bit channel is allocated a memory size equal to $LC/(mn)$. Since L_i^j is not Gaussian, we leverage the

following well-known upper bound (see, e.g. [14, Ch. 9])

$$I\left(L_i^j;\hat{L}_{i,n}^j\right) = I\left(L_i^j;L_i^j+Q_{i,n}^j\right) \leq \frac{1}{2}\log_2\left(1+\frac{\text{var}(L_i^j)}{\sigma_{i,n,j}^2}\right). \tag{26}$$

This bound allows us to obtain the conservative estimate of (i.e., upper bound on) the quantization noise power $\sigma_{i,n,j}^2$ by imposing the equality $1/2\log_2(1+\text{var}(L_i^j)/\sigma_{i,n,j}^2) = C/(mn)$, which yields

$$\sigma_{i,n,j}^2 = \frac{\text{var}(L_i^j)}{\left(2^{2C/(mn)}-1\right)}. \tag{27}$$

The variance $\text{var}(L_i^j)$ does not appear to admit a closed-form expression but it can be easily evaluated numerically. We observe that the estimate (27) is valid for the recompressed packets, i.e., for $i < n$, if the decoder employs successive refinement compression as discussed in Sec. III and Appendix.

With HARQ-CC S&C, the combined LLR for j-th bit at the n-th attempt is given by

$$\bar{L}_n^j = L_n^j + \sum_{i=1}^{n-1}\hat{L}_{i,n}^j, \tag{28}$$

hence summing the current LLRs with the previously compressed LLRs. The probability of an unsuccessful transmission for HARQ-CC S&C is finally obtained as $P_e^n = \Pr\left[\bigcap_{i=1}^n\left(\sum_{j=1}^m I(X_i(j);\bar{L}_i^j)\leq R\right)\right]$, which can be written as

$$P_e^n = \Pr\left[\bigcap_{i=1}^n\left(\frac{1}{2}\sum_{j=1}^m\sum_{b=0}^1\int f_{\bar{L}_i^j|X_i(j)}(l|b)\log_2\left(\frac{f_{\bar{L}_i^j|X_i(j)}(l|b)}{f_{\bar{L}_i^j}(l)}\right)dl\leq R\right)\right] \tag{29}$$

and evaluated similar to (22).

Remark 7. *For the same reasons explained in Remark 3, the LLR combiner (28) is optimal only when there are no HARQ buffer size limitations and it reflects the performance of a standard combiner.*

b) HARQ-CC Combine and Store (C&S): Instead of storing all the previously received LLRs, similar to Sec. III-B, HARQ-CC C&S stores the compressed value of the combined LLRs at each transmission. Specifically, if decoding of the first transmission is not successful, the stored LLR is given by

$$\hat{L}_1^j = L_1^j + Q_1^j, \tag{30}$$

where $Q_1^j \sim \mathcal{N}\left(0,\sigma_{1,j}^2\right)$ is the quantization noise. From the information-theoretic upper bound used in (26), we have $\sigma_{1,j}^2 = \text{var}(L_1^j)/\left(2^{2C/m}-1\right)$. Similar to (28), combined LLR at the n-th attempt can be written as

$$\bar{L}_n^j = L_n^j + \hat{L}_{n-1}^j \tag{31}$$

for all $m > 1$, which corresponds to the optimal combiner in the absence of quantization noise (see Remark 7). Moreover, if the n-th attempt is unsuccessful, the compressed combined LLR is given as $\hat{L}_n^j = \bar{L}_n^j + Q_n^j$, where $Q_n^j \sim \mathcal{N}(0,\sigma_{n,j}^2)$ with

quantization noise power $\sigma_{n,j}^2 = \text{var}(\bar{L}_n^j)/\left(2^{2C/m}-1\right)$, since HARQ-CC C&S allocates the available memory to store only the currently combined LLR (31). Similar to (29), the probability of an unsuccessful transmission up to the n-th retransmission is finally obtained as $P_e^n = \Pr\left[\bigcap_{i=1}^n \left(\sum_{j=1}^m I(X_i(j);\bar{L}_i^j) \le R\right)\right]$, which yields

$$P_e^n = \Pr\left[\bigcap_{i=1}^n \left(\frac{1}{2}\sum_{j=1}^m\sum_{b=0}^1 \int f_{\bar{L}_i^j|X_i(j)}(l|b)\log_2\left(\frac{f_{\bar{L}_i^j|X_i(j)}(l|b)}{f_{\bar{L}_i^j}(l)}\right)dl \le R\right)\right]. \tag{32}$$

2) HARQ-IR: With HARQ-IR, as discussed in Sec. III-C, the transmitter sends new parity bits at each transmission and the receiver stores the previously received LLRs by allocating the available memory as done for HARQ-CC S&C. Therefore, the compressed LLRs are given as (25) with (27). Moreover, using the fact that the achievable rate is the sum of all achievable rates in previously received packets [9], the probability of an unsuccessful transmission up to the n-th attempt can be calculated as $P_e^n = \Pr\left[\bigcap_{i=1}^n \left(\sum_{j=1}^m \left(I(X_i(j);L_{i,i}^j) + \sum_{k=1}^{i-1} I(X_k(j);\hat{L}_{k,i}^j)\right) \le R\right)\right]$, which yields

$$\begin{aligned} P_e^n = \Pr\Bigg[\bigcap_{i=1}^n \Bigg(&\frac{1}{2}\sum_{j=1}^m\sum_{b=0}^1 \int f_{L_{i,i}^j|X_i(j)}(l|b)\log_2\left(\frac{f_{L_{i,i}^j|X_i(j)}(l|b)}{f_{L_{i,i}^j}(l)}\right)dl \\ &+ \frac{1}{2}\sum_{k=1}^{i-1}\sum_{j=1}^m\sum_{b=0}^1 \int f_{\hat{L}_{k,i}^j|X_k(j)}(l|b)\log_2\left(\frac{f_{\hat{L}_{k,i}^j|X_k(j)}(l|b)}{f_{\hat{L}_{k,i}^j}(l)}\right)dl \le R\Bigg)\Bigg]. \end{aligned} \tag{33}$$

V. Layered Coding

So far, we have made the standard assumption that each HARQ session aims at transmitting a single data packet (carrying LR bits). Here, instead, we investigate the potential throughput gains that can be achieved via layered coding [11]. With layered coding, the transmitter encodes multiple data packets, each with a different data rate, using separate codebooks. The encoded layers are superimposed to yield the transmitted signal. Depending on the channel conditions, by the end of the HARQ session, the receiver may be able to decode only a subset of the layers. Specifically, the receiver attempts decoding starting from the first layer up to the last using a successive cancellation procedure in which higher layers are treated as noise when decoding lower layers. Layered coding is typically used to encode multimedia information sources compressed using successive refinement techniques, whereby the lower layers encode the most significant source description (see e.g., [21]). Moreover, multi-layer transmission appears to be particularly well suited to system with HARQ buffer size limitations since decoded layers can be transferred off chip and need not be retransmitted.

To elaborate, if the information rate of layer i is R_i and there are N_L layers, the throughput T can be written as

$$T = \sum_{i=1}^{N_L} \frac{R_i(1-P_{e,i}^{N_{max}})}{E[N]}, \tag{34}$$

where $P_{e,i}^n$ is the probability that layer i has not been successfully decoded up to, and including, the n-th retransmission; the

number of retransmissions N is given by (3) with

$$\Pr[N=n] = P_{e,N_L}^{n-1} - P_{e,N_L}^{n} \tag{35}$$

for $n < N_{max}$ and $\Pr[N = N_{max}] = P_{e,N_L}^{N_{max}-1}$. Note that the throughput (34) counts as useful any successfully decoded layer of information irrespective of whether, by the end of HARQ session, all the N_L layers are correctly decoded.

In the rest of this section, we study the throughput achievable with layered coding focusing, for simplicity of notation, on Gaussian signaling with baseband compression. The extension to BICM can be carried out by following the same considerations as in the previous section. Moreover, we limit the presentation to HARQ-TI and HARQ-IR. The analysis for HARQ-CC can be also performed following similar steps. Finally, similar to [11], we assume $N_L = 2$ layers[3], but the generalization to any number of layers is straightforward albeit cumbersome in terms of notation.

A. Throughput Analysis

In order to evaluate the throughput, let us define as K the random variable indicating the transmission at which the first layer is decoded correctly. Note that we have $1 \le K \le N_{max}$. In the following, we consider HARQ-IR and observe that the performance with HARQ-TI can be obtained by setting $C \to 0$ and $N_{max} = 1$ (see Remark 6).

We first fix $K = k \in \{1, \dots, N_{max}\}$ and develop the expressions for the relevant signals for the given value $K = k$. With $N_L = 2$ layers, the signal transmitted at the n-th transmission is given by the superposition

$$X_{n,k} = \begin{cases} \sqrt{\alpha \mathrm{SNR}} X_n^{(1)} + \sqrt{(1-\alpha)\mathrm{SNR}} X_n^{(2)} & \text{if } n \le k, \\ \sqrt{\mathrm{SNR}} X_n^{(2)} & \text{if } n > k, \end{cases} \tag{36}$$

where $X_n^{(i)} \sim \mathcal{CN}(0,1)$ is the signal encoding layer i at the n-th transmission and α is a power splitting factor with $0 \le \alpha \le 1$. All signals $X_n^{(i)}$ for $i = 1, 2$ and $n = 1, \dots, N_{max}$ are independent. Note that we have made the dependence of the transmitted signal on $K = k$ explicit. The received signal at the n-th transmission is given, if $K = k$, by $Y_{n,k} = H_n X_{n,k} + Z_n$, where $Z_n \sim \mathcal{CN}(0,1)$.

With HARQ-IR, at the n-th retransmission, the previously received packets are compressed and stored by allocating the available memory equally across all packets as in Sec. III and Sec. IV. As a result, the compressed i-th packet at the transmission $n \ge i$, if the first layer is decoded at the k-th transmission, is given by

$$\hat{Y}_{i,n,k} = Y_{i,k} + Q_{i,n,k}, \tag{37}$$

[3]However, unlike [11], we allow for an arbitrary number N_{max} of transmissions.

for $i = 1, \ldots, n-1$, where $Q_{i,n,k} \sim \mathcal{CN}(0, \sigma^2_{i,n,k})$ is the additive quantization noise. Similar to (8), from rate-distortion theory, the variance $\sigma^2_{i,n,k}$ can be obtained as

$$\sigma^2_{i,n,k} = \begin{cases} \left(\text{SNR}|H_i|^2 + 1\right) / \left(2^{C/n} - 1\right) & \text{if } n \neq k, \\ \left((1-\alpha)\text{SNR}|H_i|^2 + 1\right) / \left(2^{C/n} - 1\right) & \text{if } n = k. \end{cases} \tag{38}$$

Note that, for $n = k$, the first layer is removed prior to compression and hence the power of the signal to be compressed is reduced. Moreover, we observe that cancellation of the first layer does not decrease the quantization noise of the packets that have been already compressed.

The sum of the achievable rates, i.e., mutual informations, for the first layer at the transmission $n \leq k$, can be obtained as

$$R_1(n) = \log_2\left(1 + \frac{\alpha\text{SNR}|H_n|^2}{1 + (1-\alpha)\text{SNR}|H_n|^2}\right) + \sum_{i=1}^{n-1} \log_2\left(1 + \frac{\alpha\text{SNR}|H_i|^2}{1 + \sigma^2_{i,n-1,k} + (1-\alpha)\text{SNR}|H_i|^2}\right), \tag{39}$$

in which the second layer is treated as additional noise, along with the quantization nose. Note that the rate $R_1(n)$ in (39) is statistically independent of K due to the definition (38) and hence we have only emphasized the dependence on n. Similarly, the accumulated rate for the second layer at the n-th retransmission for $n \geq k$ can be written as (see also [11])

$$R_2(n,k) = \begin{cases} \log_2\left(1 + (1-\alpha)\text{SNR}|H_n|^2\right) + \sum_{i=1}^{n-1} \log_2\left(1 + \frac{(1-\alpha)\text{SNR}|H_i|^2}{1+\sigma^2_{i,n-1,k}}\right) & \text{if } n = k, \\ \log_2\left(1 + \text{SNR}|H_n|^2\right) + \sum_{i=1}^{k} \log_2\left(1 + \frac{(1-\alpha)\text{SNR}|H_i|^2}{1+\sigma^2_{i,n-1,k}}\right) & \\ \qquad + \sum_{i=k+1}^{n-1} \log_2\left(1 + \frac{\text{SNR}|H_i|^2}{1+\sigma^2_{i,n-1,k}}\right) & \text{if } n > k. \end{cases} \tag{40}$$

Note that (40) accounts for the facts that the second layer is considered for decoding only after the first layer is decoded, and that the first layer is cancelled from the received signal prior to decoding of the second layer. We also remark that $R_2(n,k)$ depends on the value $K = k$.

The probability of an unsuccessful transmission for the first layer at the n-th transmission is given by

$$P^n_{e,1} = \Pr\left[\bigcap_{i=1}^{n} \left(R_1(i) < R_1\right)\right], \tag{41}$$

where the probability is taken, here and for the rest of this section, with respect to the distribution of the channel discussed in Sec. II. The probability of an unsuccessful transmission for the second layer at the n-th transmission is given by

$$\begin{aligned} P^n_{e,2} &= \sum_{k=1}^{n} \Pr\left[K = k\right] \Pr\left[\bigcap_{j=k}^{n} \left(R_2(j,k) < R_2\right) \middle| K = k\right] \\ &= \sum_{k=1}^{n} \Pr\left[K = k\right] \prod_{j=k}^{n} q(j,k) \\ &= \sum_{k=1}^{n} \left(P^{k-1}_{e,1} - P^{k}_{e,1}\right) \prod_{j=k}^{n} q(j,k), \end{aligned} \tag{42}$$

where the first equation follows from the law of total probability and the second from the chain rule with the definition

$$q(j,k) = \Pr\left[R_2(j,k) < R_2 \middle| \bigcap_{i=1}^{k-1} \left(R_1(i) < R_1\right) \bigcap \left(R_1(k) \geq R_1\right) \bigcap_{i=k}^{j-1} \left(R_2(i,k) < R_2\right)\right]. \tag{43}$$

VI. MULTIPLE-ANTENNA LINKS

While the analysis has focused so far on single-antenna systems, in this section we elaborate on some of the additional challenges and opportunities that arise in the design of compression for HARQ buffer management when considering multiple-antenna, or MIMO, links. Specifically, we consider a MIMO link with N_t transmit antennas and N_r receive antennas. The $N_r \times 1$ received vector at each symbol of the i-th retransmission can be written as

$$\mathbf{Y}_i = \sqrt{\mathrm{SNR}}\mathbf{H}_i\mathbf{X}_i + \mathbf{Z}_i, \tag{44}$$

where SNR is the average signal to noise ratio per receive antenna; the $N_r \times N_t$ channel matrix $\mathbf{H}_i$ has unit power elements and changes independently at each retransmission; the $N_t \times 1$ vector of transmitted symbols $\mathbf{X}_i$ has unit average power, i.e., $E[\|\mathbf{X}_i\|^2] = 1$; and we have the additive noise $\mathbf{Z}_i \sim \mathcal{CN}(0, \mathbf{I})$. We focus on Gaussian signaling, by setting $\mathbf{X}_i \sim \mathcal{CN}(0, \mathbf{I}/N_t)$, on baseband compression and, for its relevance, on HARQ-IR. We also assume single-layer transmission and sufficiently large blocklengths so as to invoke standard information theoretic results. Extensions are left for future work.

As done throughput this paper, we assume that the receiver compresses and stores the packets by equally dividing the HARQ buffer. However, while in the single-antenna case, under the additive Gaussian quantization noise model, this allocation fully determines the quantization noise power, and hence the quantization strategy, with a multiple-antenna receiver a new design degree of freedom arises. Specifically, the designer can control the correlation of the additive Gaussian quantization noise across the received antennas. As discussed in, e.g., [16], [22], such correlation can be equivalently realized via a transform coding strategy, whereby the received signal is first processed by a linear transform and then independent noise is added to the elements of the resulting signal. We elaborate on this approach and on the optimization of the linear transform in the rest of this section.

If the n-th retransmission is not successful, the signal $\mathbf{Y}_i$ received at the i-th transmission is compressed$-$for the first time if $i = n$ or recompressed, by removing the current enhancement layer (Fig. 3), if $i < n-$ as

$$\hat{\mathbf{Y}}_{i,n} = \mathbf{A}_{i,n}\mathbf{Y}_i + \mathbf{Q}_i, \tag{45}$$

where $\mathbf{A}_{i,n}$ is a transform coding matrix to be calculated and $\mathbf{Q}_i \sim \mathcal{CN}(0, \mathbf{I})$ is the vector of independent Gaussian quantization noises. Note that model (45) is consistent with the assumed successive refinement strategy (see Fig. 3) only if the transforms $\mathbf{A}_{i,n}$ are selected so that the Markov chain $\mathbf{Y}_i - \hat{\mathbf{Y}}_{i,i} - \hat{\mathbf{Y}}_{i,i+1} \cdots - \hat{\mathbf{Y}}_{i,N_{max}-1}$ is preserved (see Appendix). This will be ensured by the strategy proposed below.

Assuming joint encoding across all transmission antennas (see, e.g., [23]), the achievable rate of HARQ-IR, at the n-th

attempt, can be written as the sum of the mutual informations

$$I(\mathbf{X}_n;\mathbf{Y}_n)+\sum_{i=1}^{n-1} I(\mathbf{X}_i;\hat{\mathbf{Y}}_{i,n}). \tag{46}$$

Therefore, we propose to design the transform matrix $\mathbf{A}_{i,n}$ so as to optimize (46) under the constraint that the HARQ buffer is equally allocated to all packets. Defining $\mathbf{\Omega}_{i,n}=\mathbf{A}_{i,n}^{\dagger}\mathbf{A}_{i,n}$, the optimization problem is stated as

$$\begin{aligned}\text{maximize}_{\mathbf{\Omega}_{i,n}\succeq 0}\ & I(\mathbf{X}_i;\hat{\mathbf{Y}}_{i,n})=\log\det\left(\mathbf{I}+\mathbf{\Omega}_{i,n}\left(\mathbf{I}+\frac{\text{SNR}}{N_t}\mathbf{H}_i\mathbf{H}_i^{\dagger}\right)\right)-\log\det\left(\mathbf{I}+\mathbf{\Omega}_{i,n}\right)\\ \text{s.t. }& I(\mathbf{Y}_i;\hat{\mathbf{Y}}_{i,n})=\log\det\left(\mathbf{I}+\mathbf{\Omega}_{i,n}\left(\mathbf{I}+\frac{\text{SNR}}{N_t}\mathbf{H}_i\mathbf{H}_i^{\dagger}\right)\right)\leq\frac{C}{n-1}.\end{aligned} \tag{47}$$

Following [22], given the eigenvalue decomposition $\mathbf{I}+(\text{SNR}/N_t)\mathbf{H}_i\mathbf{H}_i^{\dagger}=\mathbf{U}\text{diag}\left(\lambda_{i,1},\ldots,\lambda_{i,N_r}\right)\mathbf{U}^{\dagger}$ with unitary matrix $\mathbf{U}$ and ordered eigenvalues $\lambda_{i,1}\geq\cdots\geq\lambda_{i,N_r}$, an optimal solution is given by $\mathbf{\Omega}_{i,n}^{*}=\mathbf{U}\text{diag}\left(\alpha_{i,n,1},\ldots,\alpha_{i,n,N_r}\right)\mathbf{U}^{\dagger}$ with

$$\alpha_{i,n,l}=\left[\frac{1}{\mu_n}\left(1-\frac{1}{\lambda_{i,l}}\right)-1\right]^{+}, \tag{48}$$

where Lagrangian multiplier μ_n is selected so that the condition

$$\sum_{l=1}^{N_r}\log(1+\alpha_{i,n,l}\lambda_{i,l})=\frac{C}{n-1} \tag{49}$$

is satisfied. We observe that (48)-(49) guarantee that the gains $\alpha_{i,n,l}$ for $l=1,\ldots,N_r$ are non-increasing functions of the right-hand side of (49). This can be seen to imply the Markov chain mentioned above and hence the feasibility of successive refinement, as further elaborated in the Remark below.

Remark 8. *The transform coding compression strategy (45) under the proposed optimal design prescribes the choice of matrix* $\mathbf{A}_{i,n}$ *as*

$$\mathbf{A}_{i,n}=\textit{diag}\left(\sqrt{\alpha_{i,n,1}},\ldots,\sqrt{\alpha_{i,n,N_r}}\right)\mathbf{U}_i^{\dagger}. \tag{50}$$

This can be in practice accomplished by multiplying the received signal by the orthogonal transform matrix $\mathbf{U}_i^{\dagger}$ *and then multiplying the entries of the resulting vector by the corresponding gains* $\sqrt{\alpha_{i,n,l}}$ *prior to compression with independent unit-power quantization noises. Note that the matrix* $\mathbf{U}_i$ *is the Karhunen-Loeve transform for the received signal and hence the output vector has independent entries that can be independently quantized with no loss of optimality (see e.g., [22], [24]). We also observe that the fact that the gain* $\alpha_{i,n,l}$ *is non-increasing with respect to* $n=i,\ldots,N_{max}-1$ *for every* $l=1,\ldots,N_r$ *proves that the Markov chain* $\mathbf{Y}_i-\hat{\mathbf{Y}}_{i,i}-\hat{\mathbf{Y}}_{i,i+1}\cdots-\hat{\mathbf{Y}}_{i,N_{max}-1}$ *holds and hence successive refinement can be employed as discussed in Sec. III and detailed in the Appendix.*

With the optimal solution $\mathbf{\Omega}_{1,n}^{*},\ldots,\mathbf{\Omega}_{n-1,n}^{*}$ based on (48), the probability of an unsuccessful transmission up to the n-th

retransmission is obtained as

$$P_e^n = \Pr\left[I(\mathbf{X}_n;\mathbf{Y}_n) + \sum_{i=1}^{n-1} I(\mathbf{X}_i;\hat{\mathbf{Y}}_{i,n}) < R \right]$$
$$= \Pr\left[\log\det\left(\mathbf{I} + \frac{\mathsf{SNR}}{N_t}\mathbf{H}_i\mathbf{H}_i^\dagger\right) + \sum_{i=1}^{n-1} \log\det\left(\mathbf{I} + \mathbf{\Omega}_{i,n}^*\left(\mathbf{I} + \frac{\mathsf{SNR}}{N_t}\mathbf{H}_i\mathbf{H}_i^\dagger\right)\right) - \log\det\left(\mathbf{I} + \mathbf{\Omega}_{i,n}^*\right) < R \right], \quad (51)$$

which can be used in (2) to evaluate the throughput.

VII. Optimizing the Blocklength

In the previous sections, we made the classical assumption that the blocklength L is large enough so as to be able to invoke the asymptotic information-theoretic characterizations for achievable communication and compression rates. In this section, we instead turn to the investigation of the impact of the selection of the blocklength L on the HARQ throughput. This study is motivated by the facts that a large L generally entails a smaller probability of error and a more effective (vector) quantization but it also requires the storage of more information in the HARQ buffer. An optimal value of L is hence expected to result from the trade-off between these effects.

In order to study the impact of the blocklength L, we leverage recent information-theoretic studies on the finite-blocklength performance of channel coding [25] and source coding [26]. In this section, our approach is based on the same type of approximation proposed in [12] that are motivated by the studies [25], [27]. Furthermore, to account for the possibility to optimize the blocklength L, we consider the size of the HARQ buffer to be described by the total number of bits B_C that it can store (and not by the normalized value C). In this fashion, an increase in L does not entail a larger HARQ buffer. We define as b the total number of bits to be communicated in an HARQ session. Finally, similar to the previous section, we focus on the performance of HARQ-IR for a single-antenna link with Gaussian signaling and baseband compression, with the understanding that setting $B_C \to 0$ and $N_{max} = 1$ yields the performance of HARQ-TI.

A. Throughput Analysis

As done throughout this paper, for HARQ-IR, we assume that the receiver compresses and stores every packet by allocating an equal fraction of the HARQ buffer to all stored packets. In order to account for the effect of a finite blocklength L on the performance of the compressor, we leverage the main results in [26]. Accordingly, for a given tolerated performance ϵ_q that an optimal quantizer fails to compress a Gaussian signal with power P and a given quantization noise variance σ^2, the necessary

storage space is approximately given by [26]

$$L\left(\log_2\left(1+\frac{P}{\sigma^2}\right)+\sqrt{\frac{V_q}{L}}Q^{-1}(\epsilon_q)\right), \tag{52}$$

where the rate-dispersion factor V_q is defined as $V_q = 1/2\log_2^2 e$. Note that the term $\sqrt{V_q/L}Q^{-1}(\epsilon_q)$ measures the redundancy due to finite blocklength effects and that this redundancy increases with a smaller probability of compression error ϵ_q. Moreover, we observe that a quantizer failure can be detected by calculating the resulting distortion.

In order to apply the result (52) to the analysis of HARQ-IR, we observe the following. First, the number $N_s(n)$ of successfully compressed, and hence stored, packets prior to the $(n+1)$-th transmission is a random variable whose distribution depends on the selection of ϵ_q. Here, we assume that each packet Y_i at transmission i, in case of unsuccessful decoding, is stored with probability $1-\epsilon_q$ or discarded due to a compression failure with probability ϵ_q, independently for all $i = 1, \ldots, N_{max}$. The independence assumption follows from the independence of the signals Y_i, $i = 1, \ldots, N_{max}$. As a result, the variable $N_s(n)$ is binomial with parameter n and $1-\epsilon_q$. A second comment is that (52) applies also in the presence of successive refinement since, with Gaussian sources and mean squared error distribution, successive refinement can be optimally performed by quantizing at each layer the residual between the source and the previous coarser description, which is also a Gaussian source [15].

For a given number $N_s(n)$ of previously stored packets, if L is large enough, the transmission rate of HARQ-IR at the n-th transmission can be written as

$$R(n) = \log_2\left(1+\text{SNR}|H_n|^2\right) + \sum_{i=1}^{N_s(n-1)} \log_2\left(1+\frac{\text{SNR}|H_i|^2}{1+\sigma^2_{i,N_s(n-1)}}\right), \tag{53}$$

where $\sigma^2_{i,N_s(n-1)}$ is obtained from (52) as

$$\sigma^2_{i,N_s(n-1)} = \frac{\text{SNR}|H_i|^2+1}{2^{B_C/(LN_s(n-1))-\sqrt{V_q/L}Q^{-1}(\epsilon_q)}-1}. \tag{54}$$

Thus, by using (53) and the Gaussian approximation used in [12], inspired by [25], [27], the probability of an unsuccessful transmission up to the n-th attempt can be approximated as

$$P_e^n \approx E\left[Q\left(\frac{R(n)+1/(2L)\log_2 L - b/L}{\sqrt{V_c/L}}\right)\right], \tag{55}$$

where b is the number of information bits; the channel dispersion function V_c is defined as

$$V_c = \left(N_s(n-1)+1-\left(1+\text{SNR}|H_n|^2\right)^{-2} - \sum_{i=1}^{N_s(n-1)}\left(1+\frac{\text{SNR}|H_i|^2}{1+\sigma^2_{i,N_s(n-1)}}\right)^{-2}\right)\log_2^2 e; \tag{56}$$

and the expectation is taken over the channel and the variable $N_s(n-1)$, which are mutually independent. Using (55) in (2) yields an approximation on the achievable throughput, which will be taken here as the performance metric of interest.

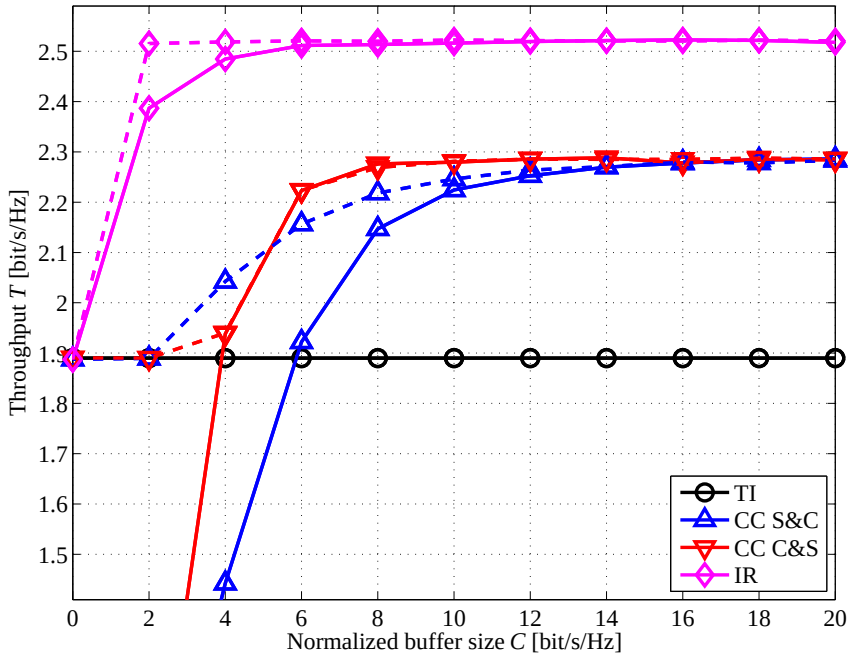

Fig. 5. Throughput T of different HARQ schemes versus the normalized buffer size C for Gaussian signaling and baseband compression without adaptive storing (solid lines) and with adaptive storing with optimal η (dashed lines) ($R = 4$ bit/s/Hz, SNR $= 10$ dB, and $N_{max} = 10$).

VIII. Numerical Results

In this section, we evaluate the throughput performance of HARQ in the presence of a finite buffer under Rayleigh fading, i.e., all channels H_i are independent zero-mean unit-power complex Gaussian variables, via numerical results. We first assume standard single-layer transmission and a large blocklength as studied in Sec. III and Sec. IV, and then we consider the impact of layered coding and of an optimized blocklength as investigated in Sec. V and Sec. VII, respectively. Lastly, we study the optimal quantization strategy for a MIMO link as proposed in Sec. VI.

We start by considering Gaussian signaling and plot the throughput of the HARQ schemes under study versus the normalized buffer size C in Fig. 5 for $R = 4$ bits/s/Hz, $N_{max} = 10$, and SNR $= 10$ dB. HARQ-IR is seen, as expected, to outperform all other strategies, but its throughput gain depends strongly on the available buffer capacity C. As for HARQ-CC, C&S is observed to be preferred over S&C, showing that the C&S mechanism uses the receiver's memory more efficiently by storing the combined packet rather than the individual packets. Moreover, the conventional Chase combiner that does not account for the impact of quantization is seen to be highly suboptimal in the regime of low C. This performance loss is recovered by implementing adaptive storing, here shown with a value of η obtained via numerical optimization for each value of C. For example, for $C = 5$ bit/s/Hz, the optimal value of η was found to be equal to 1. Note that, with adaptive storing, the S&C mechanism outperforms C&S in the low regime of C, although this is an artifact of the simple adaptive storing policy

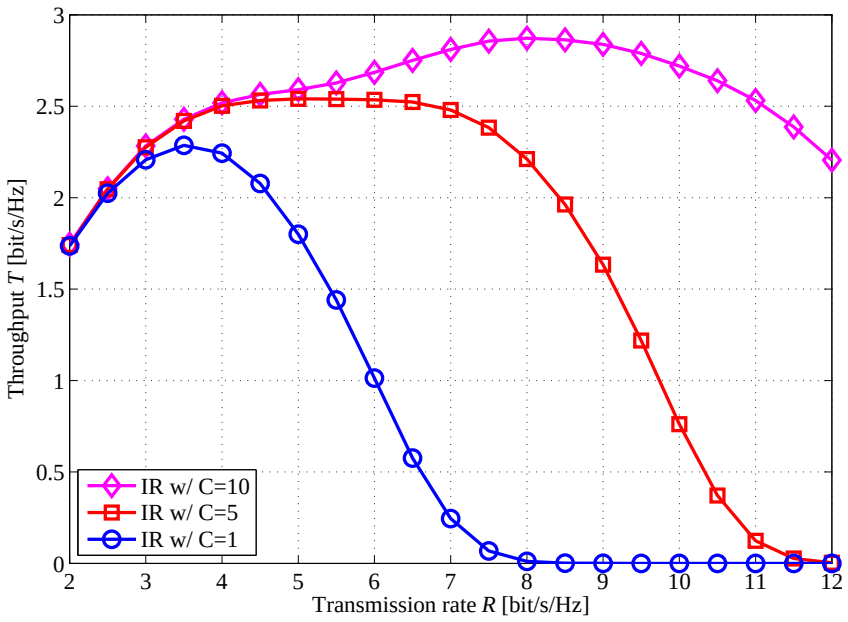

Fig. 6. Throughput T of HARQ-IR versus the transmission rate R with Gaussian signaling and baseband compression (SNR $= 10$ dB and $N_{max} = 10$).

considered here and could be fixed by implementing more sophisticated policies.

In order to illustrate the importance of accounting for the available HARQ buffer capacity when designing the HARQ strategy, as done, e.g., with limited buffer rate matching in LTE [2], [28], we plot the throughput of HARQ-IR versus the transmission rate R with SNR $= 10$ dB, $N_{max} = 10$, and different values of C in Fig. 6, It can be seen that the optimal value of R depends significantly on the value of C, ranging from around 3.5 bits/s/Hz for $C = 1$ to $R = 8$ bits/s/Hz for $C = 10$ bits/s/Hz. Further discussion on the advantages of adapting the HARQ strategy to the HARQ buffer via layered coding and blocklength optimization can be found below.

We then turn to the performance with BICM under both baseband and LLR compression. Fig. 7 and Fig. 8 show the throughput of different HARQ schemes under both compression strategies with $N_{max} = 10$ for two modulation schemes. Specifically, for Fig. 7, we set the constellation to 4-QAM, i.e., $M = 4$, and the other parameters as $R = 1.6$ bits/s/Hz and SNR $= 5$ dB; instead, for Fig. 8, we set the constellation to 16-QAM, i.e., $M = 16$, with $R = 3.4$ bits/s/Hz and SNR $= 10$ dB. Note that adaptive storing is not considered here in order to preserve the legibility of the figure but the performance with adaptive storing follows the same considerations as for Fig. 5. It is seen that baseband compression is generally advantageous over the conventional LLR compression and that the relative gain is more pronounced for simpler HARQ strategies such as TI and CC. This suggests that the use of a more sophisticated decoder, as in HARQ-IR, reduces the performance loss of a less effective compression strategy. Moreover, by comparing Fig. 7 and Fig. 8, we can see that the performance loss of LLR compression increases as the size of the constellation grows larger, particularly for simpler HARQ schemes. This is due to the

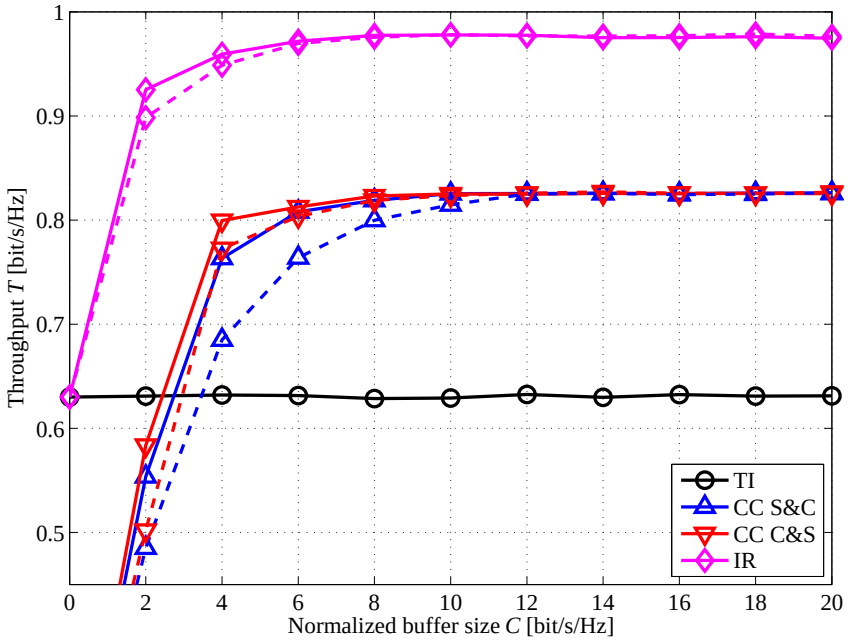

Fig. 7. Throughput T of different HARQ schemes versus the normalized buffer size C for BICM with 4-QAM for baseband compression (solid lines) and LLR compression (dashed lines) ($M = 4$, $R = 1.6$ bit/s/Hz, SNR $= 5$ dB, and $N_{max} = 10$).

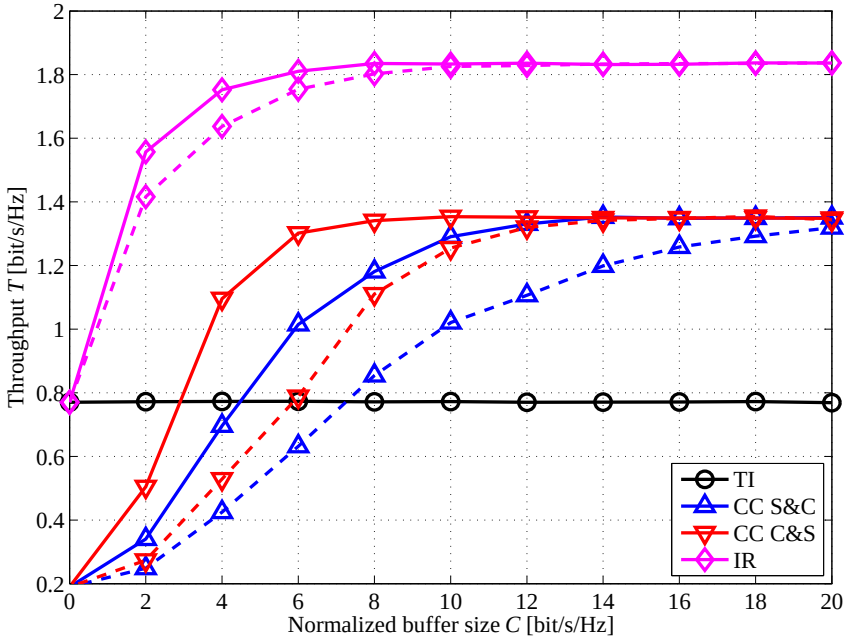

Fig. 8. Throughput T of different HARQ schemes versus the normalized buffer size C for BICM with 16-QAM for baseband compression (solid lines) and LLR compression (dashed lines) ($M = 16$, $R = 3.4$ bit/s/Hz, SNR $= 10$ dB, and $N_{max} = 10$).

larger number of LLR values that need to be compressed as the size of the constellation increases.

We now discuss the performance enhancement that can be obtained via two-level layered coding as presented in Sec. V. To this end, Fig. 9 shows the throughput of HARQ-TI and HARQ-IR with $R = 4$ bit/s/Hz, SNR $= 10$ dB, and $N_{max} = 10$.

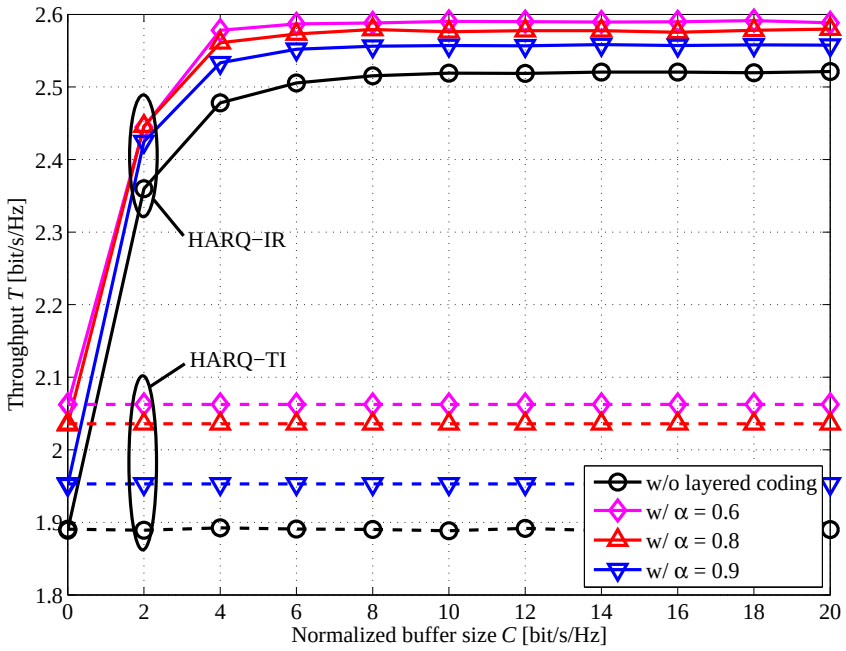

Fig. 9. Throughput T of HARQ-TI and HARQ-IR versus the normalized buffer size C for Gaussian signaling and baseband compression with layered coding with optimal R_1 ($R = 4$ bit/s/Hz, SNR $= 10$ dB, and $N_{max} = 10$).

The curves are derived by optimizing numerically over the value of the rate R_1 of the first layer with $0 \leq R_1 \leq R$, and we consider different values of the power splitting factor α. It is first observed that the throughput is quite sensitive to the choice of the power splitting factor α. Moreover, confirming the discussion in Sec. V, we see that the performance gain of layered coding is particularly pronounced in the regime of low C. For instance, the throughput is increased by 9% with layered coding at $C = 0$ bit/s/Hz with $\alpha = 0.6$, but only by 2% for a sufficiently large C.

Next, we study the throughput performance of HARQ-IR for a MIMO link in Fig. 10 following the treatment in Sec. VI. We plot the throughput gain of HARQ-IR versus the total received SNR for different numbers of transmit/receive antennas for $R = 5$ bit/s/Hz, $C = 5$ bit/s/Hz, and $N_{max} = 10$. The performance with the optimal transform coding matrix based on (48) is compared with a baseline solution in which $\mathbf{A}_{i,n} = k\mathbf{I}$, where k is selected so as to satisfy the condition (49). In Fig. 10, the throughput gain is seen to be particularly significant as the number of antenna increases and in the regime of small received SNR.

We then discuss the impact of finite blocklength in the presence of a limited HARQ buffer as per the discussion in Sec. VII. Fig. 11 shows the throughput T versus the blocklength L for different total buffer sizes B_C with $\epsilon_q = 10^{-4}$, SNR $= 5$ dB, and $N_{max} = 10$. An increase in B_C is seen to yield a significantly enhanced throughput for any value of the blocklength L, unless L is large enough to overwhelm the limited HARQ buffer. Moreover, a larger B_C calls for a reduction in the blocklength L in order to optimize the throughput. This is because, with a larger HARQ buffer, more retransmissions can be accommodated

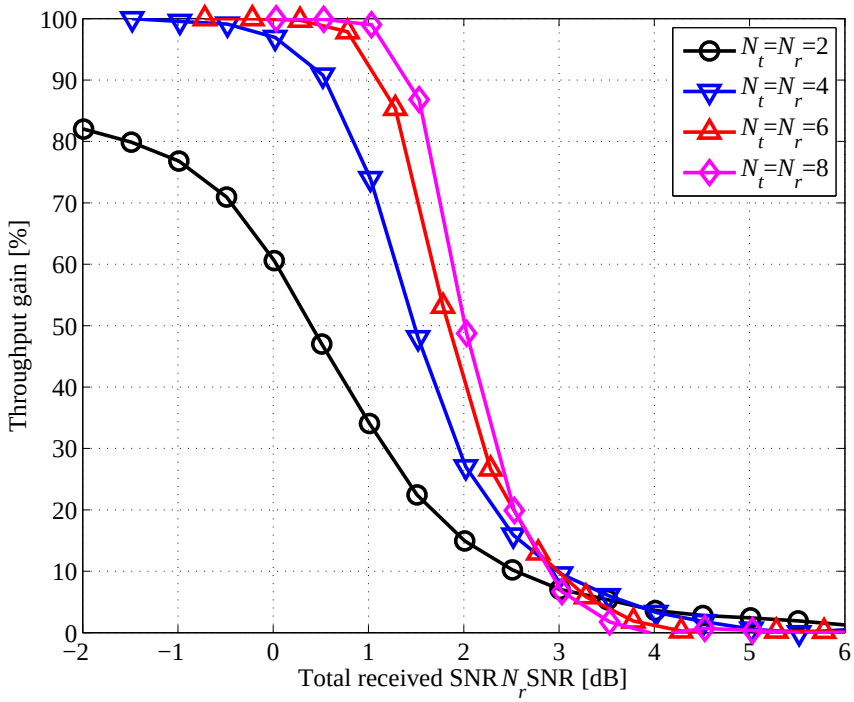

Fig. 10. Throughput gain of HARQ-IR versus the total received SNR N_rSNR for Gaussian signaling and baseband compression ($R = 5$ bit/s/Hz, $C = 5$ bit/s/Hz, and $N_{max} = 10$).

and hence it is advantageous to transmit the first packet with a more aggressive rate b/L. Finally, a smaller value of b, here $b = 1000$, yields essentially the same throughput of a larger value, here $b = 10000$, while entailing a smaller average delay. For example, for the respective throughput maximizing values of L and $B_C = 30000$, we have the average delay (see [12]) $LE[N] = 451$ with $b = 1000$ bits and $LE[N] = 4642$ with $b = 10000$ bits.

Finally, we elaborate on the effect of the compression failure probability ϵ_q in Fig. 12. We set $b = 1000$ bits, SNR $= 5$ dB, and $N_{max} = 10$. As discussed in Sec. VII, the choice of ϵ_q is one between a less significant back-off from the theoretical optimal distortion (large ϵ_q) and a smaller probability of quantization failure (small ϵ_q). For small HARQ buffers, the quantization noise is large irrespective choice of ϵ_q, and hence a small ϵ_q, which minimizes the probability of dropping received packets due to quantization errors, is to be preferred. Instead, for large HARQ buffers, the performance loss due to an excessive back-off from the optimal distortion is significant and then a larger value of ϵ_q is preferable.

IX. Conclusions

Motivated by the observation that, in modern wireless communication standards, such as LTE, the chip area occupied by the HARQ buffer is becoming increasingly significant, this work has taken an information-theoretic view of the problem of HARQ buffer management. With reference to the questions asked in the introduction, our analysis has provided three important results. (*i*) We have quantified the performance advantage that can be accrued by more sophisticated HARQ schemes such

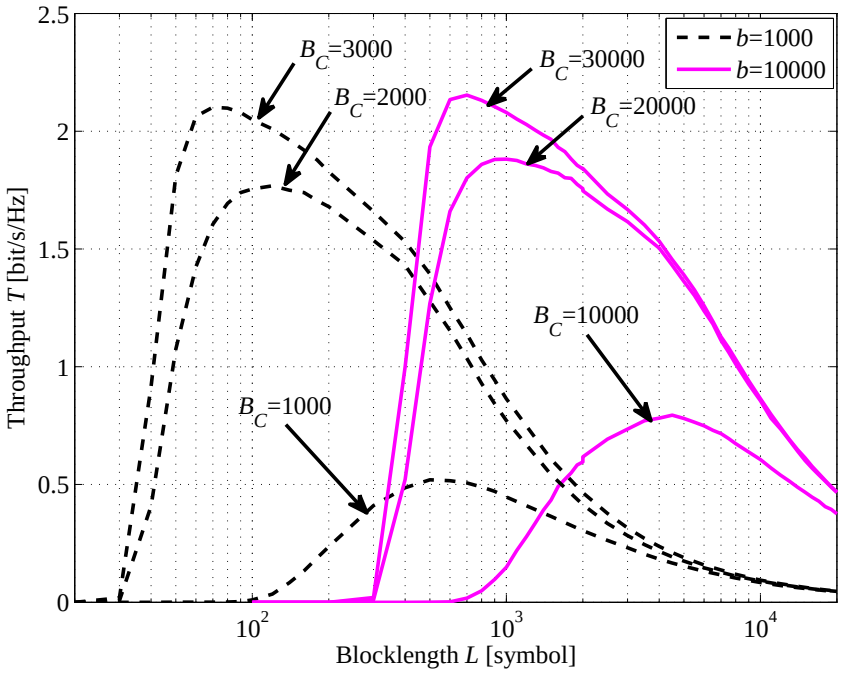

Fig. 11. Throughput T of HARQ-IR versus the blocklength L for Gaussian signaling and baseband compression ($\epsilon_q = 10^{-4}$, SNR $= 5$ dB, and $N_{max} = 10$).

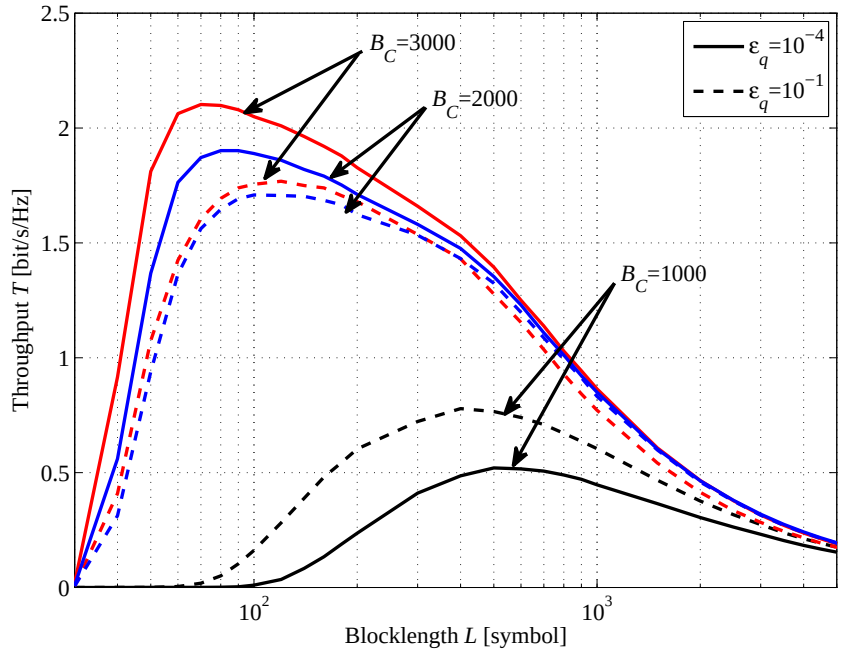

Fig. 12. Throughput T of HARQ-IR versus the blocklength L for Gaussian signaling and baseband compression ($b = 1000$ bits, SNR $= 5$ dB, and $N_{max} = 10$).

as HARQ-IR as a function of the HARQ buffer size, demonstrating that the gains depend critically on the available buffer resources. (*ii*) We have shown that storing baseband samples is generally advantageous over the conventional strategy of storing LLRs, particularly for larger constellations. Moreover, baseband compression enables sophisticated compression techniques to be implemented for multiple-antenna links, such as transform coding (see question (*iv*)). This conclusion suggests that advanced

compression mechanisms have the potential to dramatically reduce the necessary HARQ memory. (*iii*) We have investigated the potential benefits of buffer-aware transmission by considering layered modulation and the optimization of the transmission blocklength. The results demonstrate that layered modulation is particularly advantageous in the presence of small HARQ buffers, and that smaller blocklengths, and hence a more aggressive transmission rate, are beneficial for larger HARQ buffers that can accommodate more received packets.

APPENDIX

As discussed in Sec. III, the HARQ-CC S&C and HARQ-IR schemes operate by compressing all the received packets and allocating an equal fraction of the available memory to all compressed packets. Therefore, a packet that has been already compressed to $LC/(n-1)$ bits at the $(n-1)$-th transmission needs to be recompressed at the n-th transmission (if unsuccessful) to a smaller number LC/n of bits. In this section, we explain how this can be accomplished by using successive refinement (or layered) coding. In so doing, we demonstrate that the equality (7) is valid also for recompressed packets. Note that we discuss here the case of Gaussian signaling, but the treatment of BICM follows in a similar fashion.

Consider the compression of a received i-th packet as in (1). The packet can be recompressed at most $N_{max}-i$ times since there are at most as many possible retransmissions in which the packet at hand can be reused by the decoder. To enable this, if the i-th transmission is unsuccessful, the decoder compresses (1) by using a successive refinement code with $N_{max}-i$ layers, which corresponds to the progressively less accurate compressions that are stored in subsequent retransmissions. Specifically, for each packet i, we have the descriptions $\hat{Y}_{i,n} = Y_i + Q_{i,n}$ in (6) for $n = i, i+1, \ldots, N_{max}-1$. The corresponding quantization noise variances $\sigma^2_{i,n}$ are increasing in n, since the allocated memory becomes smaller in later transmissions, i.e., we have the inequalities

$$\sigma^2_{i,i} \leq \sigma^2_{i,i+1} \leq \cdots \leq \sigma^2_{i,N_{max}-1}. \tag{57}$$

As summarized in Fig. 3 for each packet i, the decoder produces the "base layer" description $\hat{Y}_{i,N_{max}-1}$, which has the largest quantization noise variance $\sigma^2_{i,N_{max}-1}$, and the "refinement layers" $\hat{Y}_{i,N_{max}-2},\ \hat{Y}_{i,N_{max}-3}, \ldots,\ \hat{Y}_{i,i}$ with progressively smaller quantization noise variances as per (57). At the j-th retransmission, with $j = i, \ldots, N_{max}-1$, only the descriptions $\hat{Y}_{i,N_{max}-1},\ \hat{Y}_{i,N_{max}-2}, \ldots,\ \hat{Y}_{i,j}$ are stored and the higher refinement layers are discarded.

Based on the discussion above, we can write the quantization noises as

$$Q_{i,n} = \sum_{j=i}^{n} \Delta Q_{i,j}, \tag{58}$$

where the variables $\Delta Q_{i,j} \sim CN(0, \sigma^2_{i,j} - \sigma^2_{i,j-1})$ are independent and represent the increase in quantization noise variance in going from the $(j-1)$-th description to the j-th (we set $\sigma^2_{i,i-1} = 0$). This shows that the Markov chain $Y_i - \hat{Y}_{i,i} - \hat{Y}_{i,i+1} \cdots -$

$\hat{Y}_{i,N_{max}-1}$ holds. Moreover, using standard information-theoretic results on the performance of successive refinement (see, e.g., [15, Ch. 13]), we obtain that the number $R_{i,n}$ of bits per symbol needed to store the n-th description is given by

$$R_{i,n} = I(Y_i; Y_i + Q_{i,n} | Y_i + Q_{i,n+1}) \tag{59}$$

for $n = i, \ldots, N_{max} - 2$ and $R_{i,N_{max}-1} = I(Y_i; Y_i + Q_{i,N_{max}-1})$. The overall number of bits per symbol that need to be stored at the n-th transmission (if unsuccessful) is hence given by

$$\sum_{j=n}^{N_{max}-1} R_{i,j} = I(Y_i; Y_i + Q_{i,n}), \tag{60}$$

which can be seen by recalling the definition of condition mutual information[4] and noticing that, because of the mentioned Markov chain relationship, we have $h(Y_i|Y_i + Q_{i,n}, Y_i + Q_{i,n+1}) = h(Y_i|Y_i + Q_{i,n})$. We conclude that the equality (7) holds also for recompressed packets.

REFERENCES

[1] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*. Prentice-Hall, 2010.

[2] S. Sesia, I. Toufik, and M. Baker, *LTE: the UMTS long term evolution*. Wiley Online Library, 2009.

[3] S. Lin and D. J. Costello Jr., *Error control coding: Fundamentals and applications*. Englewood Cliffs, Prentice-Hall, 1983.

[4] D. Chase, "Code combining-A maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Trans. Commun.*, vol. COM-33, pp. 385-393, May 1985.

[5] D. Bai, C. Park, J. Lee, H. Nguyen, J. Singh, A. Gupta, Z. Pi, T. Kim, C. Lim, M.-G. Kim, and I. Kang, "LTE-advanced modem design: Challenges and persepectives," *IEEE Commun. Magazine*, vol. 50, no. 2, pp. 178-186, Feb. 2102.

[6] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. of the IEEE*, vol. 102, no. 3, pp. 366-385, Mar. 2014.

[7] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellural: It will work!," *IEEE Access*, pp. 335-349, May 2013.

[8] M. Danieli, S, Forchhammer, J. D. Andersen, L. P. B. Christensen, and S. S. Christensen, "Maximum mutual information vector quantization of Log-Likelihood Ratios for memory efficient HARQ implementations," in *Proc. Data Compression Confer. (DCC 2010)*, pp. 30-39, Mar. 2010.

[9] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1971-1988, Jul. 2001.

[10] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 927-946, May 1998.

[11] A. Steiner and S. Shamai (Shitz), "Multi-layer broadcasting hybrid-ARQ Strategies for block fading channels," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2640-2650, Jul. 2008.

[12] R. Devassy, G. Durisi, P. Popovski, and E. G. Ström, "Finite-blocklength analysis of the ARQ-protocol throughput over the Gaussian collision channel," *Int. Symposium Commun., Control, Signal Process. (ISCCSP)*, pp. 173-177, May 2014.

[13] W. Lee, O. Simeone, J. Kang, R. Sundeep, and P. Popovski, "HARQ buffer management: An information-theoretic view," *IEEE Int. Symposium Inform. Theory (ISIT)*, submitted to, Jun. 2015.

[4] $I(A; B|C) = h(A|C) - h(A|C, B)$ for continuous jointly distributed variables A, B, C, where $h(A)$ is the differential entropy of variable A.

[14] T. M. Cover and J. A. Thomas, *Element of information theory*. John Wiley&Sons, 2006.

[15] A El. Gamal and Y. H. Kim, *Network information theory*. Cambridge University Press, 2011.

[16] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inform. Theory* vol. 42, no. 4, pp. 1152-1159, Jul. 1996.

[17] V. Nagpal, I.-H. Wang, M. Jorgovanovic, D. Tse, and B. Nikolic, "Coding and system design for quantize-map-and-forward relaying," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 8, pp. 1423-1435, Aug. 2013.

[18] M. W. Marcellin and T. R. Fischer, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Trans. Commun.*, vol. 38, no. 1, pp. 82-93, Jan. 1990.

[19] A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1520-1529, Sep. 1996.

[20] S. Rosati, S. Tomasin, M. Butussi, and B. Rimoldi, "LLR compression for BICM systems using large constellations," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2864-2875, Jul. 2013.

[21] C. T. K. Ng, D. Gunduz, A. J. Goldsmith, and E. Erkip, "Distortion minimization in Gaussian layered broadcast coding with successive refinement," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5074-5086, Nov. 2009.

[22] A. Del Coso and S. Simoens, "Distributed compression for MIMO coordinated networks with a backhaul constraint," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4698-4709, Sep. 2009.

[23] H. Zheng, A. Lozano, and M. Haleem, "Multiple ARQ processes for MIMO systems," *EURASIP J. App. Sig. Proc.*, vol 5, pp. 772-782, 2004.

[24] S.-H. Park, O. Simeone, O. Sahin and S. Shamai (Shitz), "Robust Layered Transmission and Compression for Distributed Uplink Reception in Cloud Radio Access Networks, *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 204-216, Jan. 2014.

[25] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2307-2359, May 2010.

[26] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 58, no. 6, pp. 3309-3338, Jun. 2012.

[27] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inform. Theory*, vol. 60, no. 7, pp. 4232-4265, Jul. 2014.

[28] J.-F. Cheng, A. Nimbalker, Y. Blankenship, B. Classon, and T. K. Blankenship, "Analysis of circular buffer rate matching for LTE turbo code," in *Proc. IEEE Veh. Technol. Confer. (VTC)*, pp. 1-5, Sep. 2008.